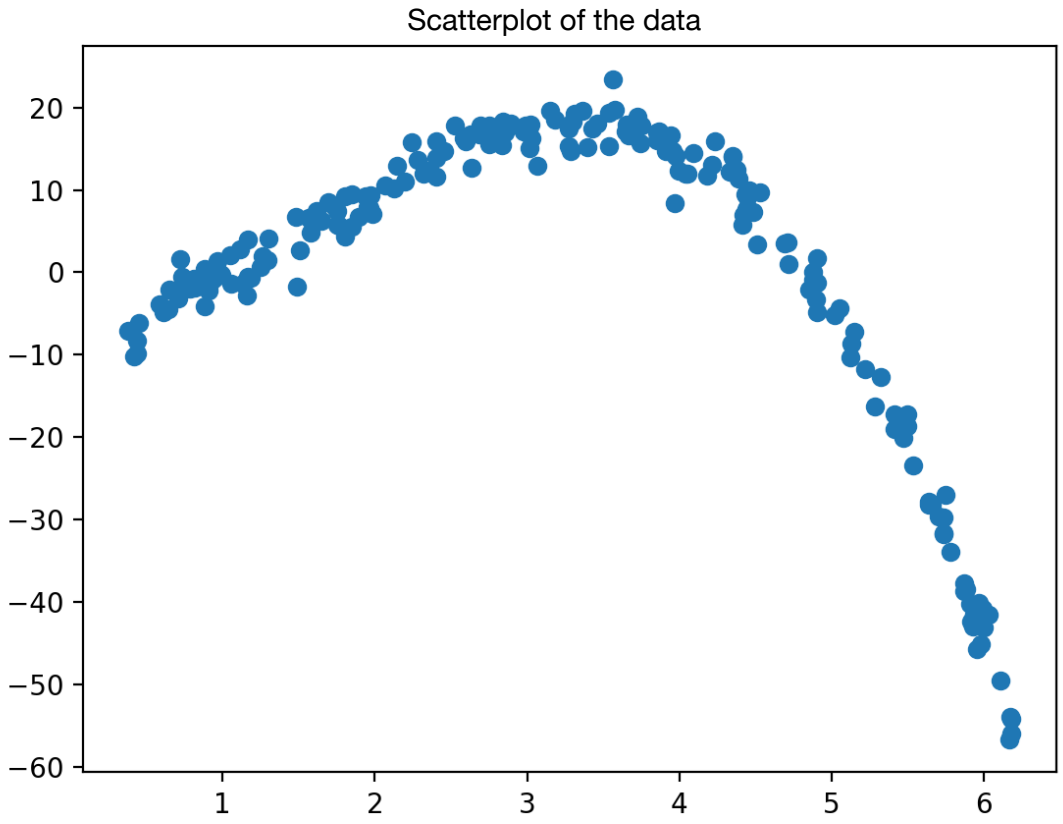
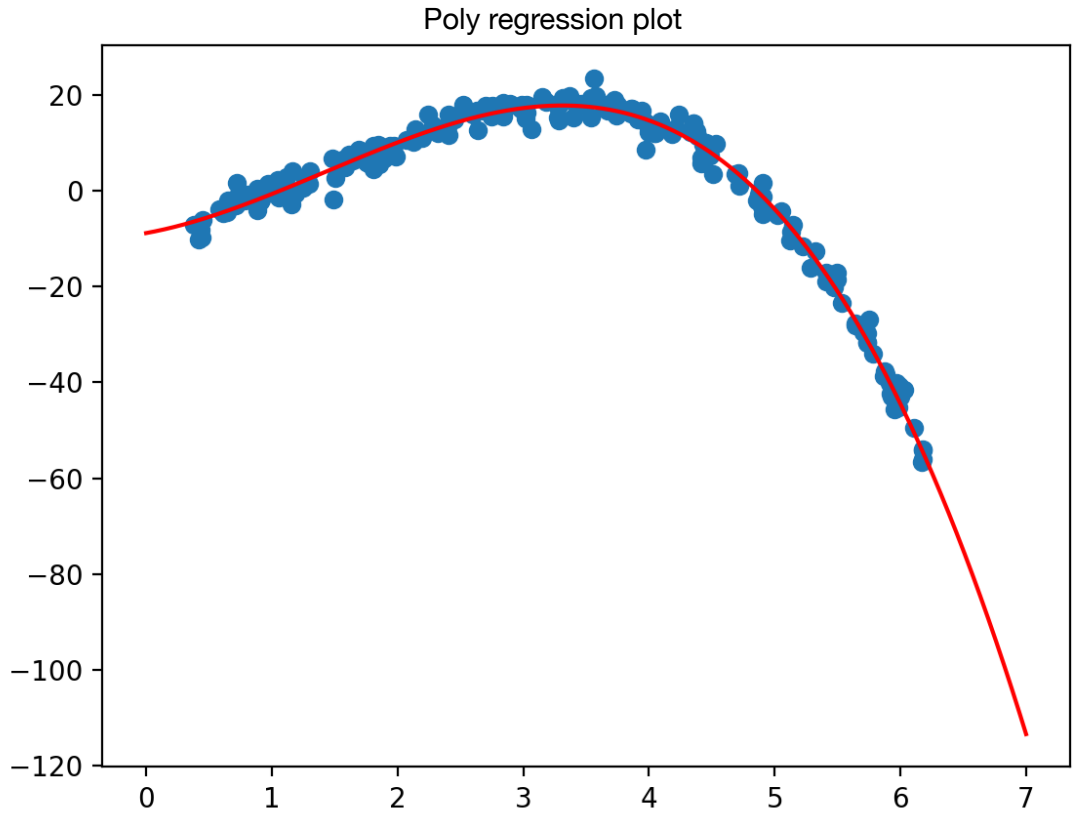
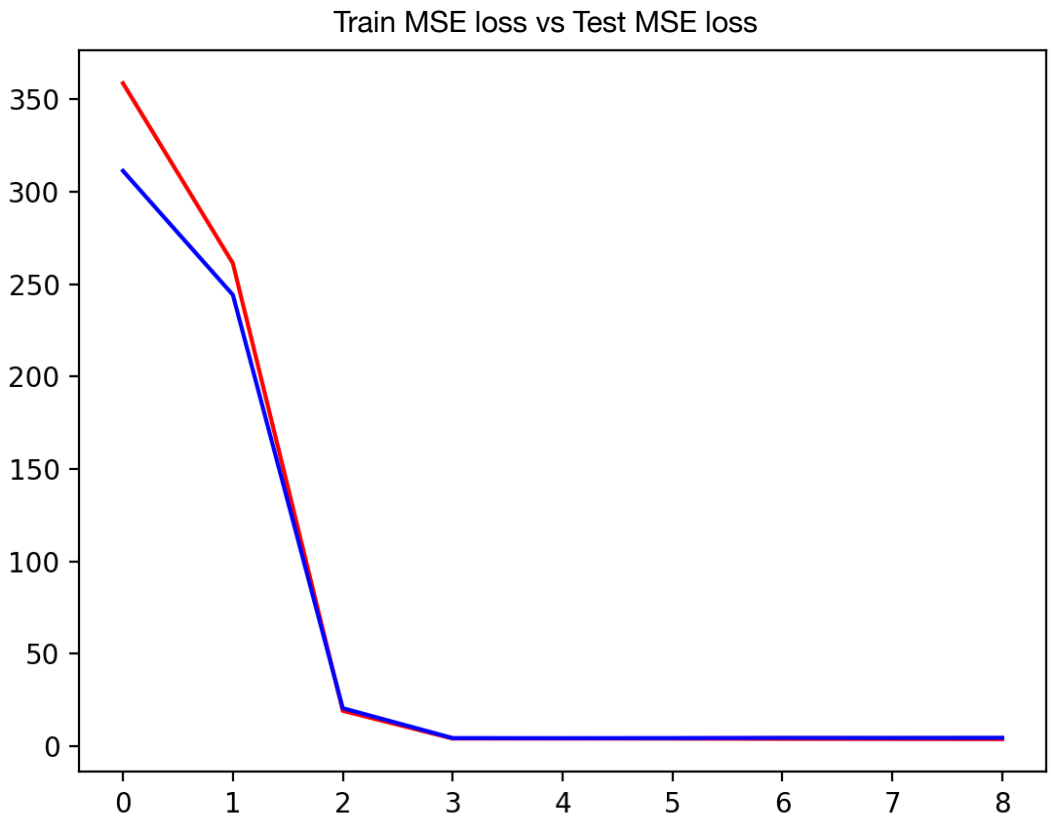
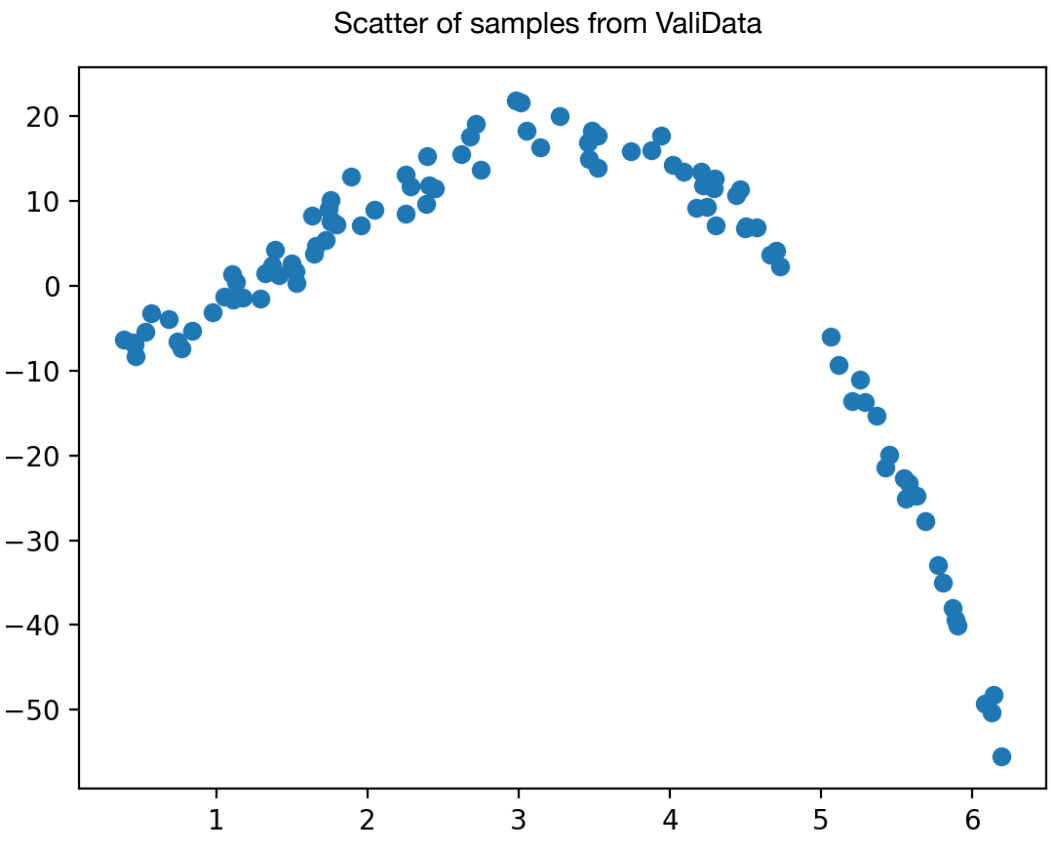
Q1- Polynomial Regression Model Fitting

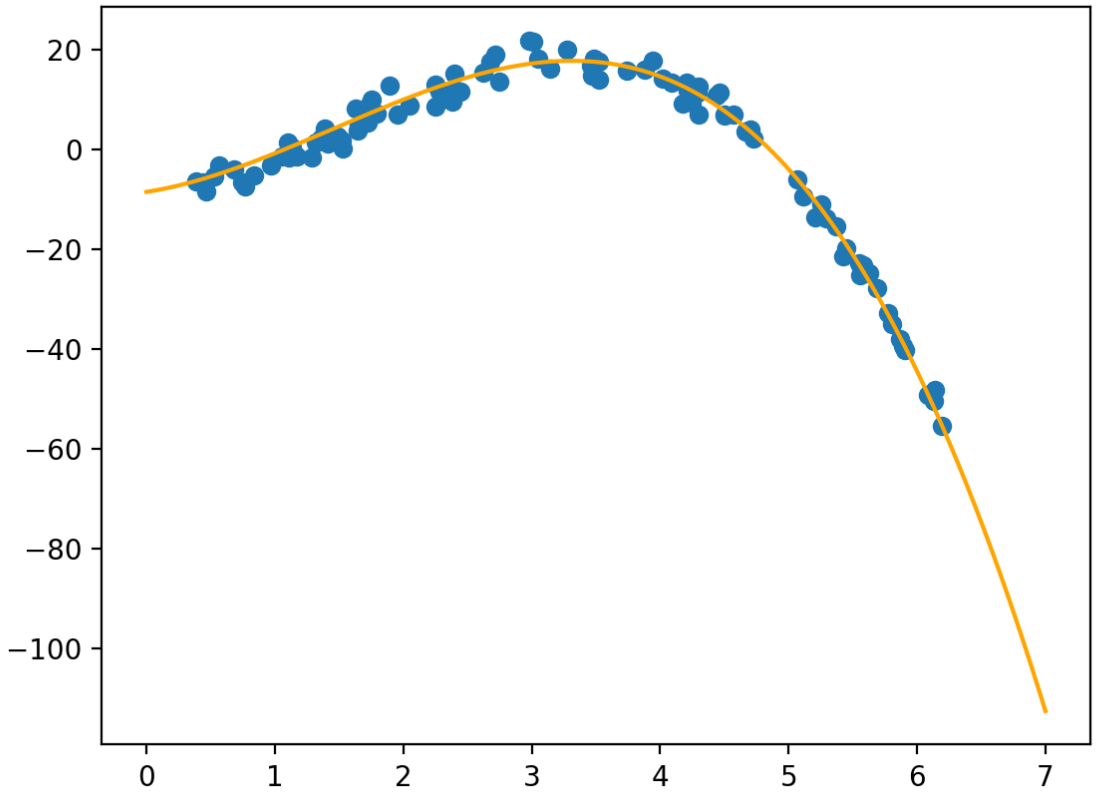





Train MSE loss vs Test MSE loss
350
300
250
200
150
100
50
0
0
1
2
3
4
5
6
7
8


Scatter of samples from ValiData
20
10
0
−10
−20
−30
−40
−50
1
2
3
4
5
6

Best-fit polynomial curve obtained using pthetaBest




```
Poly Theta:  [-8.87149489  4.79000177  4.3909502  -1.02976757]
MSE:  4.071469897324788
MSE Test Loss:  4.3319184830212745
```

```
Train Validate Poly Regression Output
Min train loss:  3.75601057666721
Min test loss:  4.290269812941981
Train loss when test loss is min:  4.069409721278246
Poly of order:  4
Best poly order if using train data:  8
Test loss when train loss is min:  4.504321206273911
```

To generate the best theta, we break up our data into a testing and training data sets and use the training data set to learn the best theta for a given order. We then record the mean training error for each order (using the testing data set) and choose the theta that gives us the smallest error.

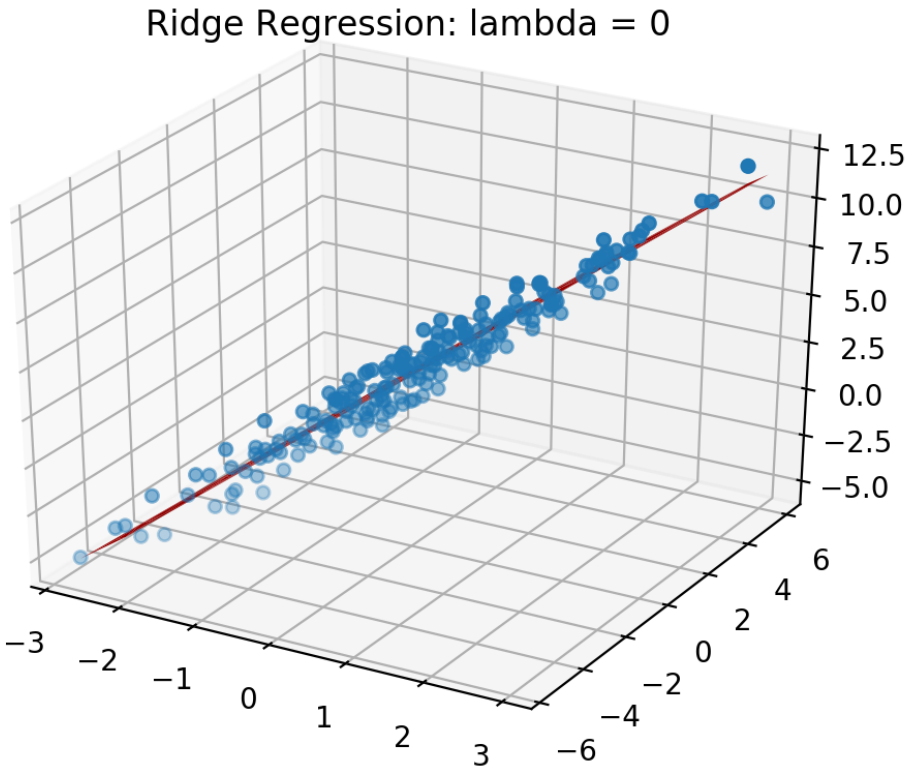
Q2 - Ridge Regression

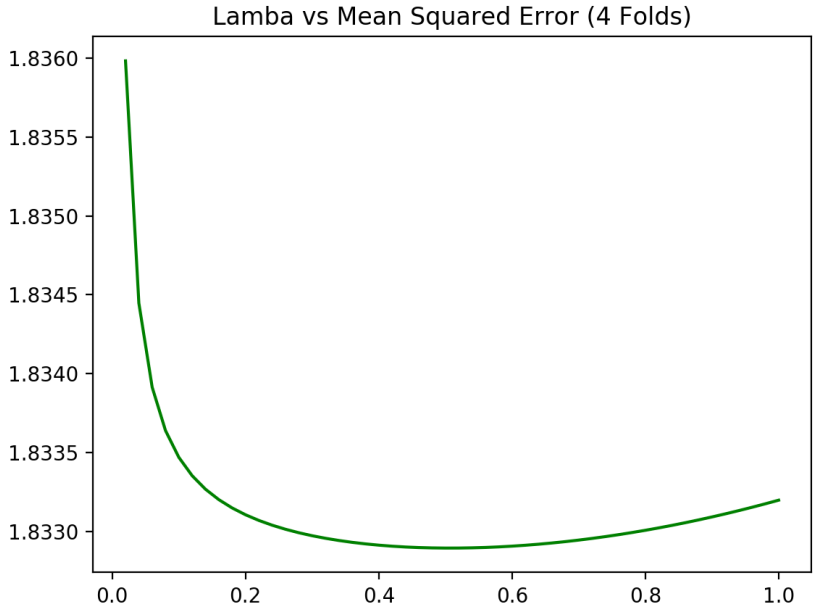
1.1) We have that (Y−Xβ)^T*(Y−Xβ) is the sum of the squares of the residuals. We then add the penalty term λβ^T*β and then derive on β to get the normal equation, similar to how linear regression gets its normal equation because in both cases we are minimizing loss which requires deriving. Once we have the normal equation X^T*Y=(X^T*X+λI)*β, you can solve for the ridge regression estimator.

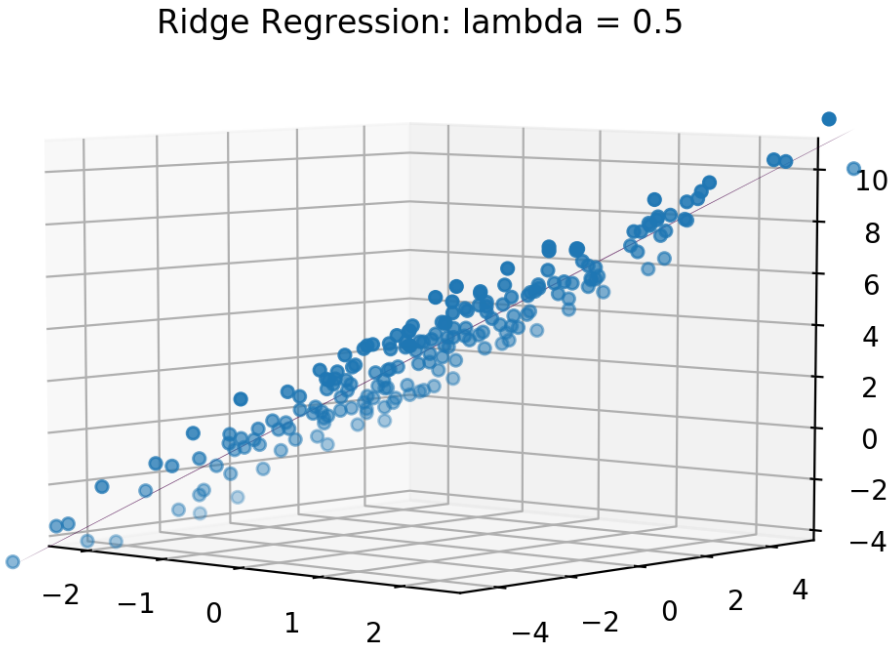
1.2) We have a 3x2 and a 3x1 matrix. We could solve this through linear regression because we have to do X^T * Y in the normal equation – transposing X gives us a 2 x 3 matrix, and a 2 x 3 and 3 x 1 matrix can be multiplied successfully without giving a dimension error.
However, since x1 and x2 are linearly dependent, where 2*x1 = x2, we could choose one of the features (x1 or x2) and perform linear regression only using that feature.

1.3) If you know beforehand that the coefficient should be sparse, you should use lasso regression since it introduces sparsity into the model.

1.4)



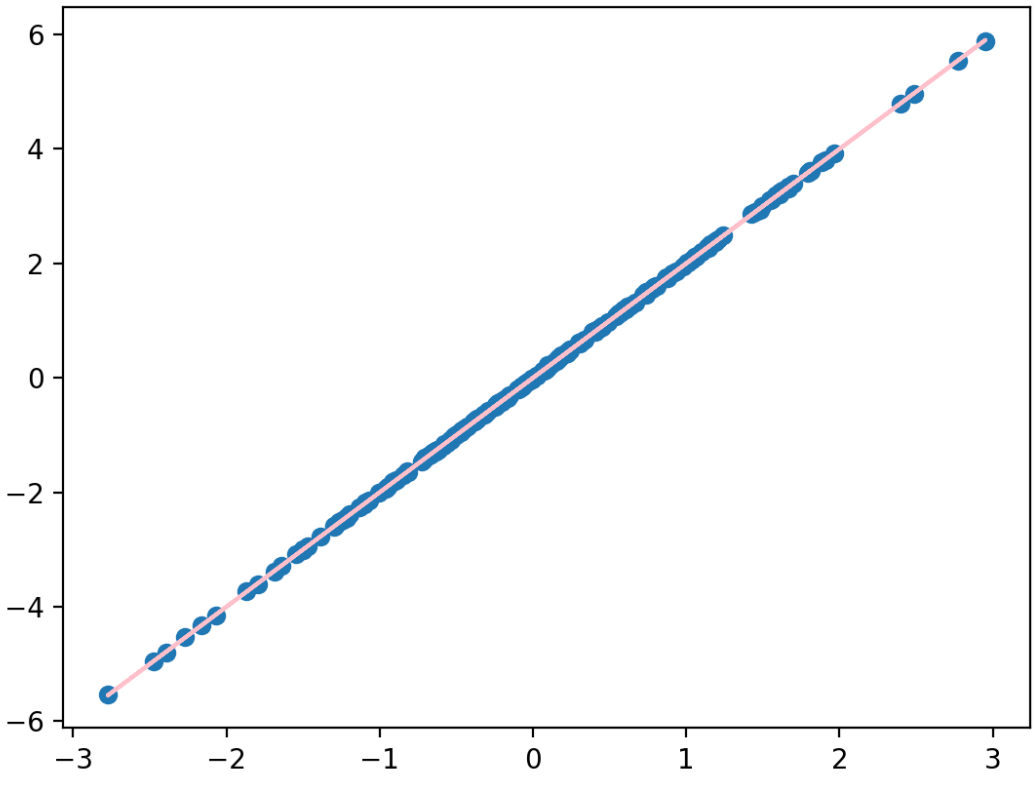





```
betaLR:
[  2.97139801 -11.00332214   6.96229098]
lambdaBest:  0.5
betaRR:  [2.9655605  0.48050468 1.22043384]
```

1.5) Below is the LR plot of X1 and X2 from the ridge regression dataset.

```
Theta:  [2.66735631e-04 2.00013256e+00]
```



Since the two features given in the data set (X1 and X2) are linearly dependent, performing LR on the data set is bound to cause issues, considering that the data in each column is not providing unique feature information. Thus Ridge Regression would perform better on the given data.

Q3 - Sample Exam Questions

- (True/False). Ridge regression model increases the bias but reduces the variance comparing to the linear regression model.

  True, ridge regression shrinks the estimated coefficients towards zero which reduces the variance, and it has more bias than linear regression which has really low bias.

- (True or False?) The error of a model measured over its training set provides a pessimistically biased estimate of the true error of the model.

  False, the training error is optimistically biased, not pessimistically, since in most cases this value is smaller than the actual error.

- (True or False?) If you are given m data points, and use half for training and half for testing, the difference between training error and test error decreases as m increases.

  True, because as the number of data points increases, both the training error and test error will converge to the actual error value.

- (True or False?) Overfitting is more likely when the set of training data is small.

  True, because with a small set of training data, you can get a hypothesis that fits the hypothesis too closely more easily than with a larger set, which results in overfitting.

- (True or False?) Overfitting is more likely when the model's hypothesis space is small.

  False, since small hypotheses spaces result in less variance and more bias. This makes it more difficult to get a hypothesis that fits the hypothesis too closely, meaning that overfitting is not more likely with a small hypothesis space.

- (True/False) When the tuning parameter λ increases its value, the parameter β in the ridge regression will not converge to zero vector, since Lasso enforces sparsity on β (assuming no bias term here).

False, because the ridge regression parameter (estimator) will still converge to zero even with the sparsity being enforced, since it will need to decrease to minimize the effect of the tuning parameter increasing.

- (True/False). Ridge regression can fit the data well even if its feature variables have certain linearly dependent relationships among each other.

  True, because compared to other estimates, ridge regression estimates are more stable, and aren't as affected by such relationships since it reduces the magnitude of the relationships between independent variables in the model.

Question 2.

No, these aren't very useful basis functions to use, because for the set, only inputs 1, 3, or 5 will ever produce an output, so possible values such as 2 and 4 will cause problems, and it won't fit the data very well for these either (y will always be zero when x is 2 or 4).

Question 3.

The mean square LOOCV error in this case would be zero, since all three points on shown in the figure are on the line y = x^2 which fits the model exactly. Thus regardless of which point we leave out, the error will still be zero.